CLAIMS

What is claimed is:

1. A processor-based method, comprising:

selecting a set number of functions correlating variable parameters of a dataset; and

clustering the dataset by iteratively applying a regression algorithm and a K-Harmonic Means performance function on the set number of functions.

2. The processor-based method of claim 1, wherein said clustering comprises:

determining distances between datapoints of the dataset and values correlated with the set number of functions;

regressing the set number of functions using datapoint probability and weighting factors associated with the determined distances;

calculating a difference of harmonic averages for the distances determined prior to and subsequent to said regressing; and

repeating said regressing, determining and calculating upon determining the difference of harmonic averages is greater than a predetermined value.

3. The processor-based method of claim 2, wherein said determining the distances comprises determining distances from each datapoint of the dataset to values within each function of the set number of functions.

4. The processor-based method of claim 2, wherein said selecting and said clustering are conducted for a plurality of datasets each from a different data source.

5. The processor-based method of claim 4, wherein said selecting and said clustering are conducted in parallel for each of the plurality of datasets.

6. The processor-based method of claim 4, further comprising determining a common coefficient vector to compensate for variations between similar sets of functions within the different data sources.

7. The processor-based method of claim 6, wherein said determining the common coefficient vector comprises:

developing matrices from the dataset datapoints and the probability and weighting factors for each of the datasets prior to said reiterating; and

determining the common coefficient vector from a composite of the developed matrices.

8. The processor-based method of claim 7, further comprising multiplying the similar sets of functions within the different data sources by the common coefficient vector.

9. A storage medium comprising program instructions executable by a processor for:

selecting a set number of functions correlating variable parameters of a dataset;

determining distances between datapoints of the dataset and values correlated with the set number of functions;

calculating harmonic averages of the distances;

regressing the set number of functions using datapoint probability and weighting factors associated with the determined distances;

repeating said determining and calculating for the regressed set of functions;

computing a change in harmonic averages for the set number of functions prior to and subsequent to said regressing; and

reiterating said regressing, repeating and computing upon determining the change in harmonic averages is greater than a predetermined value.

10. The storage medium of claim 9, wherein the program instructions are executable using a processor for computing the datapoint probability and weighting factors.

11. The storage medium of claim 9, wherein the program instructions are executable using a processor for developing matrices from the dataset datapoints and the probability and weighting factors prior to said reiterating.

12. The storage medium of claim 11, wherein the program instructions are executable using a processor for amassing matrices developed from a plurality of datasets each from a different data source.

13. The storage medium of claim 11, wherein the program instructions are executable using a processor for determining a common coefficient vector from the composite of matrices.

14. The method of claim 13, wherein the program instructions are executable using a processor for multiplying similar sets of functions within the different data sources by the common coefficient vector.

15. A system, comprising:
an input port configured to receive data; and
a processor configured to:
regress functions correlating variable parameters of a set of the data;
cluster the functions using a K-Harmonic Mean performance function; and
repeat said regress and cluster sequentially.

16. The system of claim 15, wherein the processor is arranged within one of a plurality of data sources each comprising a processor configured to:
regress the functions on a dataset of the respective data source;

cluster the functions using a K-Harmonic Mean performance function; and

repeat said regress and cluster sequentially.

17. The system of claim 15, further comprising a central station coupled to the plurality of data sources, wherein the central station comprises a processor configured to compute common coefficient vectors which compensate for variations between the regressively clustered functions representing the datasets, and wherein each of the processors of the data sources is configured to alter the functions by the common coefficient vectors.

18. A system, comprising:

a plurality of data sources; and

a means for regressively clustering datapoints from the plurality of data sources without transferring data between the plurality of data sources.

19. The system of claim 18, wherein the means for regressively clustering the datasets comprises a means for applying a regression algorithm and a K-Harmonic Means performance function on the datasets.

20. The system of claim 18, wherein the means for regressively clustering the datasets comprises a means for applying a regression algorithm and a K-Means performance function on the datasets.

21. The system of claim 18, wherein the means for regressively clustering the datasets comprises a means for applying a regression algorithm and an Expectation Maximization performance function on the datasets.

22. The system of claim 18, further comprising a central station communicably coupled to the plurality of data sources, wherein the means is further for:

collecting dataset information at the central station from the plurality of data sources;

determining a common coefficient vector from the collected dataset information; and

altering datasets within the plurality of data sources by the common coefficient vector.

23. The system of claim 18, wherein the means for regressively clustering the datasets comprises a storage medium with program instructions executable using a processor for:

selecting a set number of functions correlating variable parameters of a dataset;

determining distances between datapoints of the dataset and values correlated with the set number of functions;

regressing the set number of functions using datapoint probability and weighting factors associated with the determined distances;

calculating a difference of harmonic averages for the distances determined prior to and subsequent to said regressing; and

reiterating said regressing, determining and calculating upon determining the difference of harmonic averages is less than a predetermined value.

24. A system, comprising:

a plurality of data sources each having a processor configured to access datapoints within the respective data source; and

a central station coupled to the plurality of data sources and comprising a processor, wherein the processors of the central station and plurality of data sources are collectively configured to mine the datapoints of the data sources as a whole without transferring all of the datapoints between the data sources and the central station.

25. The system of claim 24, wherein the each of the processors within the plurality of data sources is configured to regressively cluster a dataset within the respective data source.

26. The system of claim 25, wherein the processor within the central station is configured to:

collect information pertaining to the regressively clustered datasets; and

based upon the collected information, calculate common coefficient vectors which balance variations between functions correlating similar variable parameters of the regressively clustered datasets.

27. The system of claim 26, wherein the processor within the central station is further configured to:

compute a residual error from the common coefficient vectors;

propagate the common coefficient vectors to the data sources upon computing a residual error value greater than a predetermined value; and

send a message to the data sources to terminate the regression clustering of the datasets upon computing a residual error value less than a predetermined value.

28. A processor-based method for mining data, comprising:

independently applying a regression clustering algorithm to a plurality of distributed datasets;

developing matrices from probability and weighting factors computed from the regression clustering algorithm, wherein the matrices individually represent the distributed datasets without including all datapoints within the datasets;

determining global coefficient vectors from a composite of the matrices; and

multiplying functions correlating similar variable parameters of the distributed datasets by the global coefficient vectors.

29. The processor-based method of claim 28, further comprising repeating said independently applying, said developing, said determining and said multiplying.

30. The processor-based method of claim 28, further comprising calculating a residue error associated with the global coefficients prior to said multiplying.